Quantitative Methods

Human Sciences
Coursework assignment, 2020–2021

In this assignment, you will analyse variation in age-standardised all-cause mortality rates across approximately 3,000 counties in the United States. To get started, upload the two data files named `county_mortality.csv` and `county_covariates.csv` into R. The first file, drawn from the National Center for Health Statistics, contains the following variables:

- ⋆ `county`: name of county.
- ⋆ `fips`: unique five-digit Federal Information Processing Standards (FIPS) code assigned to each county.
- ⋆ `state`: name of the state in which the county is located.
- ⋆ `all_cause`: measure of age-standardised mortality rates from all causes per 100,000 county population.

The second file, drawn from the US Census Bureau, contains the following variables:

- ⋆ `countyname`: name of county.
- ⋆ `FIPS`: unique five-digit Federal Information Processing Standards (FIPS) code assigned to each county.
- ⋆ `statename`: name of the state in which the county is located.
- ⋆ `income_median`: county-level median household income, measured in constant US dollars.
- ⋆ `edu_hs`: fraction of county population with at least a high school diploma.
- ⋆ `race_black`: fraction of county population who are African American.
- ⋆ `race_other`: fraction of county population who belong to some other non-White ethnic minority.
- ⋆ `pop_density`: county-level population density.

1. Create a single data set that is ready for further analysis by merging these two files.

2. Create a suitable visualisation of how all-cause mortality rates vary across counties.

3. Specify a linear regression model of the form $Y = X_1\beta_1 + X_2\beta_2 + \cdots + \epsilon$ in which you regress all-cause mortality (Y) on all or some of the other variables ($X_1, X_2, \ldots$) and where ϵ is an error term. Explain what assumptions this model is making and justify your choice of covariates (including why you have or have not included a given variable). Then present your key findings. Include a discussion and interpretation of parameter estimates, model uncertainty, and model fit.

4. Suppose we are interested in isolating the net association between median household income (X_1) and all-cause mortality (Y). Run the model $Y = \tilde{X}_1\beta_1 + \epsilon$, where $\tilde{X}_1$ is the residual from a regression of X_1 on all the other covariates. In other words, first run a model in which you regress median household income on all the other covariates, save the residuals, and then regress all-cause mortality on these residuals. Compare the parameter estimate of β_1 in this model to the estimate of β_1 in your previous model. What do you notice? Can you provide an intuitive explanation?

5. Provide a suitable visualisation that conveys how all-cause mortality rates vary by median household income, after adjusting for the other covariates.

6. Note that the impact of one additional dollar of household income on the number of deaths per 100,000 population may be relatively small and not always easy to interpret. Express your key quantity of interest — i.e., the net association between income and mortality — on some more easily interpretable scale (e.g., in terms of percentages, standard deviations, quantiles, or some other suitable measure which you would find easy to interpret and convey to a general readership).

7. Discuss the strengths and weaknesses of your model(s). What conclusions can you plausibly draw from your analysis?

Present your analysis in the form of a brief report (approximately 5 pages, including tables and figures) and append a clearly annotated and replicable R script that shows your work. Alternatively, you may integrate the R code into the main body of the report to illustrate each step of your analysis.

Good luck!